DOCUMENT RESUME

ED 419 842 TM 028 418

AUTHOR Haley, Kathleen
TITLE Watkins-Farnum Revisited: Application of Modern Test Theory to Music Performance Assessment.
PUB DATE 1998-04-00
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Difficulty Level; Evaluation Methods; Intermediate Grades; Junior High Schools; Middle Schools; *Music; *Performance Based Assessment; Pilot Projects; Scaling; *Test Theory
IDENTIFIERS Hierarchical Models; Middle School Students; *Rasch Model

ABSTRACT

A study was proposed to determine to what extent a hierarchical structure exists in music and in tests used to measure music ability. The first research question was whether items in the Watkins-Farnum Performance Scale (J. Watkins and S. Farnum, 1954) (WFPS) form a hierarchy, so that early exercises (bars played) are generally easier than later ones. Another question was whether the items in the Clarinet Performance Rating Scale (CPRS) (H. Abeles, 1973) form a hierarchy of difficulty, and a third was the relationship between WFPS and CPRS scores. The application of the Rasch model to these measures of musical performance was explored. Rasch scaling would offer advantages in terms of ability estimates for music students taking the WFPS. Data from approximately 245 students for the WFPS, collected for another study, are to be used in the projected study, along with data from 50 clarinet players who will have scores on both the WFPS and the CPRS. In a pilot study involving 125 middle school students, Rasch analysis was applied to WFPS results, and data were, in general, a good fit to the model. These results suggest the suitability of the Rasch model for the proposed study. (Contains 4 figures and 12 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 419 842

Watkins-Farnum Revisited: Application of Modern Test Theory to Music Performance Assessment

Kathleen Haley

This paper is prepared for the:
Annual Meeting of the American Educational Research Association in San Diego, CA
April 1998

TM028418

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Kathleen Haley
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

The music educator has a particularly challenging job in assessing his or her students' individual performances. In many ways, it can be compared to judging the quality of an English essay. For instance, there are technical issues: sentences must be structured according to grammatical rules; performers must play the correct notes and rhythms. There also is an element of confounding, in which a difficulty in one area makes it more difficult to express one's ability in another area. For example, an English student may have imaginative ideas, but be constrained by a physical difficulty in writing, or poor grammar and sentence structure. Similarly, a poor tone quality or lack of breath support can cause all manners of difficulty in articulation, flexibility, intonation, and almost every area of playing an instrument. Finally, there are subjective issues, issues of performing or writing style and interpretation.

One divergence that separates the music teacher from the English teacher and from teachers of the visual arts is that music exists temporally, while the visual arts, and to some extent written expression, exist spatially. It is a simple matter to reexamine certain sections of an essay or a work of art, but this is not so with a musical performance. Only modern technology makes any reexamination possible, and still no review of a section in keeping with the context of the whole is possible. This further complicates the assessment process. This fact, combined with the subjective and philosophical issues of aesthetics, makes the task of the music educator or adjudicator that much more difficult. My dissertation research is designed to

learn to what extent a hierarchical structure exists in music and in tests used to measure musical ability. This paper will discuss the context of the study, describe the data collection and analysis procedures, and report the results of a pilot study. The following questions are being asked.

Research Questions

1) Do the items in the Watkins-Farnum Performance Scale (Watkins & Farnum, 1954) form a hierarchy, as theorized by Watkins, such that bars[1] in exercise one are generally easier than those in exercise two, and so forth?

2) Do the items within the Clarinet Performance Scale form any hierarchy such that certain musical characteristics are more difficult to achieve? Within each subscale, do the items form a hierarchy?

3) Given that the Watkins-Farnum Performance Scale (WFPS) is a standardized instrument which does not consider style or interpretation, and that the Clarinet Performance Rating Scale (CPRS) is non-standardized and attempts to consider a musical performance in a more subjective way, what is the strength of the relationship between student scores on the WFPS and on the CPRS?

[1] This topic contains numerous opportunities for misunderstood words. The word "measure" is both a unit of musical notation and an assessment tool. I will attempt to use the equivalent word "bar" to denote the musical meaning of the word. Furthermore, the word "scale" has both a measurement meaning and a musical meaning. Only the measurement meaning is used in this paper.

Background

There are a number of ways in which modern test theory can advance the understanding of assessment in music education. This paper will particularly consider the Rasch model, one of a class of item response theory models, and how it can be applied to two measures of musical performance.

The Rasch Model

The Rasch model is a method of analyzing test results such that not only are examinees given ability estimates, but test items are also given difficulty estimates. The two sets of estimates are on the same scale and can be compared. That is, if a student of average ability attempted an item of below-average difficulty, he or she would be more likely than not to answer it correctly. The Rasch model can be applied to many types of data. It will be explained in this and the next chapter in general terms. However, in the context of this paper, to "answer an item correctly" means to play a bar correctly.

Another advantage of the Rasch model is that differences in the difficulty of questions is taken into account when computing an estimate for an individual. That is, an examinee who takes an easy test and answers some wrong and some right would *not* be expected to obtain the same estimate as one who took a more difficult test and answered the same proportion correctly. The examinee who answered the more difficult questions would receive the higher estimate. Two examinees of equal ability would be expected to receive the same estimate, regardless of the items they took.

Similarly, two items of the same difficulty would receive the same difficulty estimate, whether they were administered to high- or low-ability examinees.

Roughly speaking, we may divide the research into music performance assessment into two strands, standardized and non-standardized. In this sense, standardized will refer to tests in which each examinee performs the same music *and* administration and scoring procedures are the same for each examinee. Non-standardized will refer to those tests in which one or more of these conditions is not met, such as when students select their own music but are all assessed using the same process. This dissertation will explore the use of the Rasch model to music assessment, and discuss how it may be applied to one test from each strand of research. The two examples are discussed, followed by a brief discussion of the Rasch model and how it may be applied to each.

Standardized assessment

In 1942, John G. Watkins published his landmark dissertation, *Objective Measurement of Instrumental Performance*, in which he developed a highly valid and reliable scale for the measurement of cornet performance. It consists of fourteen short exercises of progressive difficulty. Each bar of each exercise is scored correct if there are no errors, and incorrect if one or more errors are made. Dr. Stephen E. Farnum later modified the scale to be useful for any instrument (Watkins & Farnum, 1954). The result is the Watkins-Farnum Performance Scale (WFPS), a scale which has been in

widespread use by instrumental music teachers ever since. However, there were two things that Dr. Watkins wanted to do that classical test theory did not provide the technology to do. With item response theory, these are now possible.

1) Since the scale is composed of exercises that become progressively more difficult, Watkins stopped his subjects once "a sheer chaos of sound in no manner resembling music was coming forth from the horn." This practice prevented him from being able to calculate difficulty values for each bar, since one could not assume that a student would not play a single bar correctly in the exercises not reached. Instead, he calculated the probability of playing a given exercise with a specified number of errors. Using item response theory, difficulties can be calculated even if all subjects in a sample do not attempt the item.

2) In an effort to save time, Watkins wrote a preliminary exercise which began quite simply, but became difficult quickly. The intent was to give test administrators an idea of the playing level of the subject, so that advanced players would not have to play the easiest exercises. The preliminary exercise was dropped because of low reliability. However, once the items are calibrated under the Rasch model, the additional information we will have about each item will allow us to compute an ability estimate using only one or a few exercises, provided the student does not play them perfectly or completely fail them. Thus the scale will have the built-in time saving feature Watkins had hoped to create.

Nonstandardized assessment

In 1973, Harold Abeles published the Clarinet Performance Rating Scale (CPRS). This was a scale created through factor analysis of statements about clarinet performances. The scale consists of thirty items, including six subscales of five items, each rated on a five point Likert scale. It can be used to rate a performance of any piece of music, as opposed to the above measure, in which music is provided. Abeles reported high reliability and validity coefficients. Later studies have replicated his work for trombone (Kidd, 1981), band (DCamp, 1980), and chorus (Cooksey, 1977).

For the CPRS, no structure is presupposed. The Rasch calibration would therefore be exploratory, in order to determine if any hierarchy exists among types of statements about musical performances. The six subscales are interpretation, tone, rhythm, intonation, tempo, and articulation. Application of the Rasch model would allow interpretation as to whether a hierarchy exists, both among the subscales and within each subscale.

Significance of the study

For administration of the Watkins-Farnum Performance Scale

In its current form, the WFPS can be cumbersome to administer, particularly to more advanced students. An examiner must hear and score all exercises beginning with the first, up to the point at which the student is completely unable to perform the exercises. By scaling under the Rasch model, a quasi-adaptive version can be constructed, which would allow an

instructor to select an appropriate starting point, based on his or her knowledge of the student, and estimate an ability level based on fewer exercises.

Another benefit of Rasch scaling is that, if the items fit the model, a student would be expected to correctly perform most items below a certain level and not perform most items above the same level. With this knowledge, a set of examples could be constructed to give an examinee with a given ability estimate an idea of the type of music he or she should be practicing in order to improve.

For knowledge of the hierarchy of musical skills

Since the CPRS has six subscales, it is of interest to find whether each is of approximately equal difficulty for students. One of many reasons this is important is that early music instruction often focuses on technical skills, particularly rhythm and pitch. As students advance, they are taught more about style and expression. If the categories are generally of equal difficulty, this removes much of the justification for this sequence of teaching.

For the validity of the Watkins-Farnum Performance Scale

The WFPS is often criticized for measuring only objective aspects of musical performance. This exclusion of subjective aspects was intentional on Watkins' part in an attempt to increase the reliability. However, this narrow focus does raise important concerns about validity. Since the CPRS focuses on a broader set of musical qualities, its correlation with scores on the WFPS

is of interest. A high correlation would indicate that, although the WFPS does not measure subjective aspects of performance, those who have high scores on the WFPS also tend to score highly on the more subjective aspects. This would provide evidence of the validity of the WFPS as an *indirect* measure of overall musical performance.

Limitations of the Study

One difficulty that may reasonably be troubling the reader is the idea that the WFPS is objective, measuring only the presence or absence of errors, while the quality of a musical performance is a subjective decision. Indeed, this is a well-founded concern and one which receives little treatment in Watkins dissertation. However, as will be discussed more thoroughly in the next chapter, the subjective qualities of musical performance, often specified as "expression," "musicality," or similar labels, are highly correlated with global ratings and with the more objective qualities measured in the WFPS. Nevertheless, it is not recommended that a music teacher use the WFPS, either with or without the Rasch item response scaling that will be developed in this study, for decisions that affect the future of an individual student. That is, this test should not, except in small part, determine a student's grade or be used to make a decision whether or not to admit a student to a performing ensemble. Some appropriate uses would be:

1) A first hearing of new students, to allow a director to become better acquainted with a student and his or her ability.

2) An initial seating of students within a section, given that there will be future opportunity for students to move up within the section.

3) A monitor of progress in specific areas, such as rhythm or articulation, or in sightreading.

4) A valid measure of musical performance to be used to report group scores in future research studies.

While this study touches broad issues in music assessment, its primary focus is on assessment that can be used in the classroom. Therefore, it has not been possible to make use of some of the most recent trends in testing, particularly those that require multiple raters. However, the use of the Multi-faceted Rasch model (Linacre, 1989) to adjust scores for differences in rater severity would be a promising avenue of future research.

Methods

Instruments

The Watkins-Farnum Performance Scale

The Watkins-Farnum Performance Scale is a set of 14 exercises of increasing difficulty, ranging from very simple to quite difficult. Each exercise is sixteen to 36 bars long. Two versions in the same format exist, Form A and Form B; however, this paper will consider only Form A. Examinees play the exercises in order, either after practice or at sight. The adjudicator notes each bar in which an error occurs. Only one error is scored in each bar. Therefore, the possible scores for each bar are one and zero only. The maximum possible

score on each exercise is a given standard, and the total points scored equals the standard for the exercise minus the number of bars containing an error. A student is to be allowed to continue until he/she scores zero on two consecutive exercises. The total score for the test is the sum of the individual exercise scores. Exercises not reached are scored zero under the assumption that the probability of successfully playing an exercise after completely failing two exercises of lower difficulty is extremely small.

The Clarinet Performance Rating Scale

The Clarinet Performance Rating Scale was developed in 1973 by Harold Abeles. It consists of 30 Likert-type items that are descriptive of a clarinet performance. Adjudicators are asked to "Strongly agree," "Agree," "Neither agree nor disagree," "Disagree," or "Strongly disagree" with descriptors such as "Effective musical communication," or "Flat in low register." The 30 items are organized into six subscales.

Data Collection Procedures

Each subject will be tape recorded performing the WFPS. The purpose of the tape recorder is twofold. The first is to eliminate the nervousness caused when a performer can see an adjudicator making notes during a performance. The second purpose is to allow future study with these data, including an inter-rater reliability study. Research has shown that judgments of performance ability may be made equally reliably whether live, videotaped, or audiotaped (Vasil, 1973; Massel, 1978), and regardless of the quality of the recording (Vasil, 1973).

Each performance will be scored twice. The investigator will score each performance through the WFPS protocol described earlier in this chapter. Only one hearing of each performance will take place in this scoring process since this will be the situation for most users of the test. Each bar will be scored correct or incorrect for each subject, resulting in 132 "items" per subject. The sixth exercise will then be scored using the CPRS protocol. This scoring will be done by the investigator and by one other adjudicator, and the percent agreement calculated. The sixth exercise was chosen because it is complex enough that the performer will have to make stylistic decisions, but easy enough that most performers will reach that exercise.

Data Analysis

This research involves two samples of students. The first is a set of existing data, with scores on the WFPS only. This sample consists of approximately 245 students, who play any band instrument. The second sample, yet to be collected, will consist of approximately 50 clarinet players and will have scores on both the WFPS and the CPRS. Research question 1 will be answered using the two samples combined, while questions 2 and 3 will be answered using the second sample only.

Research Question 1: Do the items in the Watkins-Farnum Performance Scale (Watkins & Farnum, 1954) form a hierarchy, as theorized by Watkins, such that bars in exercise one are generally easier than those in exercise two, and so forth?

Quest Interactive Test Analysis System (Adams & Khoo, 1993) will be used to calibrate the scale. That is, each bar of the WFPS will receive a difficulty estimate, and each student will receive an ability estimate. It is to be expected that bars from the first exercise will have lower difficulty estimates than those from the second exercise, and that a similar pattern should be evident through all eight exercises. Item fit statistics will be examined to determine the extent to which the items collectively meet the requirements of the Rasch model.

Research Question 2: Do the items within the Clarinet Performance Scale form any hierarchy such that certain musical characteristics are more difficult to achieve? Within each subscale, do the items form a hierarchy?

The CPRS will be calibrated in the same manner as the WFPS, using the sample of students with scores on both the WFPS and the CPRS. While no particular item ordering is expected, difficulty estimates will be inspected to determine whether items within a subscale tend to have similar estimates. Item fit statistics will be examined to determine the extent to which the items collectively meet the requirements of the Rasch model.

Research Question 3: Given that the Watkins-Farnum Performance Scale (WFPS) is a standardized instrument which does not consider style or interpretation, and that the Clarinet Performance Rating Scale (CPRS) is non-standardized and attempts to consider a musical performance in a more subjective way, what is the strength of the relationship between student scores on the WFPS and on the CPRS?

For the sample of students with scores on both the WFPS and the CPRS, the Pearson correlation will be computed between student ability estimates on the two scales. Given the high validity coefficients for the CPRS, the correlation may be considered a concurrent validity coefficient for the WFPS.

Pilot Study

The pilot study addresses research question one, and is an analysis of existing data from the Watkins-Farnum Performance Scale. The sample consists of 125 students from a Rhode Island middle school. The students were administered the test at the end of their second year of instrumental instruction. A Rasch analysis was performed using Quest Interactive Test Analysis System (Adams & Khoo). For purposes of the analysis, total score was to equal the number of bars played correctly. Two items were deleted, the first item in the first exercise due to a perfect score, and the first item in the seventh exercise because of a zero score. A fit analysis was also performed.

Results

The resulting variable map generally conformed to Watkin's theory about the relative difficulty of the exercises. Higher numbered items refer to more difficult exercises and are generally located higher in the scale than the lower numbered items. In the variable map that follows, items are numbered on the right side, with the more difficult items toward the top. Students are designated as X's on the left side, with the more able students toward the top.

Items are separated by exercise, with those items from exercise 1 in the leftmost column, and those from exercise 8 in the column furthest right.

BEST COPY AVAILABLE

It was expected that items within an exercise would cluster together, but that there should be no particular order within an exercise. The exercises and their respective items are as follows:

Exercise 1	1-16	(item 1 deleted)
Exercise 2	17-32	
Exercise 3	33-48	
Exercise 4	49-64	
Exercise 5	65-80	
Exercise 6	81-96	
Exercise 7	97-116	(item 97 deleted)
Exercise 8	117-132	

That is, the highest items on the scale should be numbers 117 through 132, in no particular order. Numbers 98 through 116 should generally be below these, but above 81-96. There are many exceptions to the general trend of items belonging to difficult exercises being difficult themselves. This is not surprising since within a difficult exercise one would not be surprised to find isolated bars which low-performing students are able to play correctly. Note also that the items from exercise 1 and exercise 2 are perhaps more intermingled on the map than any other pair of adjacent exercises. This is due to the fact that almost all students were able to play most of these bars, and therefore there is little information to distinguish the difficulty of exercise 2 from that of exercise 1. If more very low performing students were

added, some would be expected to correctly play bars from exercise 1 but not those from exercise 2, thus separating the items.

The data in general were quite a good fit to the model. Among the seven items showing substantial misfit, there were two items that showed severe misfit This means that higher-scoring students are not necessarily playing the bar correctly at a higher rate than low-scoring students. One of the items, item 64, was an easy bar with a repeat sign in it. If the student missed the repeat, the bar was marked wrong. This item might be better scored correct if no other error was made, since the repeat sign is not actually part of the bar but a direction to begin again. The other item, item 54, showed no readily interpretable reason for the misfit. The following table shows each misfitting item along with its associated difficulty estimate and fit statistic. For a table including all 132 items, see appendix D. The infit t is a measure of fit that is approximately distributed as t. Items with a |t| value greater than 2 are included in this table.

Item	Difficulty estimate	Infit t
34	-0.67	2.6
38	-0.67	3.3
46	-0.55	3.6
54	1.03	5
55	0.81	2.1
64	0.26	5.4
77	0.48	2.9
84	0.43	-2.9
86	0.37	-2.6
88	-0.12	-4.1
96	-0.39	-3.4
100	1.09	-5.2

Item	Difficulty estimate	Infit t
102	0.75	-4.8
104	0.86	-4
106	0.75	-5.2
107	1.9	-3.3
108	0.92	-4.8
109	1.09	-3.6
110	2.28	-2
112	1.66	-2.5
114	2.28	-2
116	1.31	-3.6
126	3.36	-2.2
132	3.36	-2.3

Quite a few items were flagged for overfit. Since the model generates *probabilities* that a student will perform an item correctly, we would not expect every student with a probability over .5 to respond correctly to that item. As this deterministic condition is approached, the item is considered to "overfit." Seventeen items overfit the model to some degree; four items overfit severely. However, all of these items were in the last three exercises, which the lowest-scoring students were not given the opportunity to play. The overfit is therefore expected because all of the lowest scoring students got these items "wrong." If enough high scoring students were added to calibrate all fourteen exercises in the scale, we would expect this problem to disappear, but to reappear in the very hardest exercises.

Discussion

When Watkins created the scale, he intentionally wrote exercises of increasing difficulty. He verified this judgment by asking experts to rate the

difficulty of the exercises. In pilot testing the instrument, he was further able to test this assumption. He was only able to test this at the exercise level, because different numbers of students took each item. With Rasch scaling, difficulty values can now be calculated for each item. This is important because, as mentioned, the difficulties of the exercises are rough; many easy bars are found within difficult exercises, and vice versa. However, the two most important contributions of the Rasch scaling are efficiency and information quality.

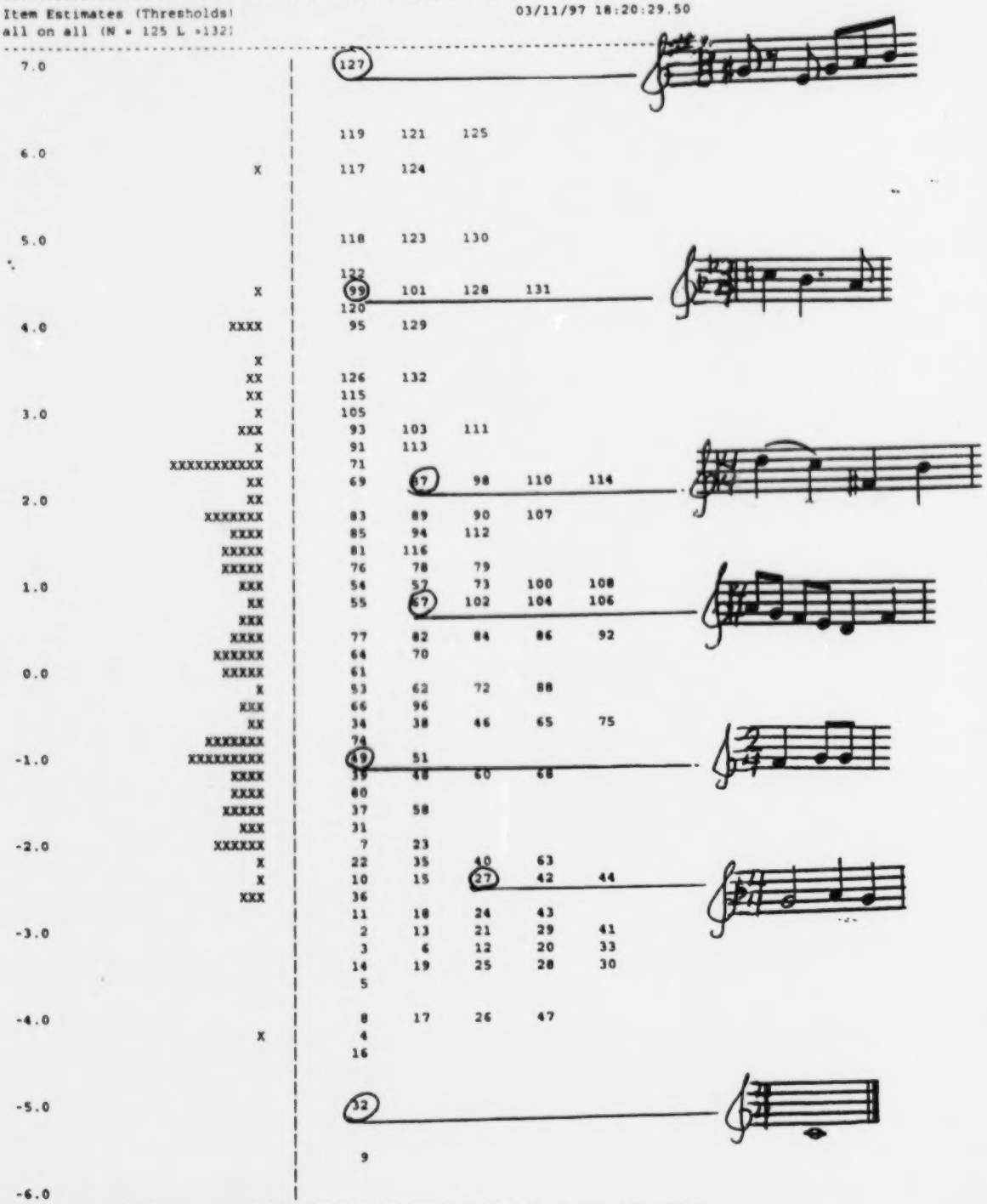
The ordinary method of administering the WFPS is to have the student begin at exercise one, and perform each exercise in turn. The scorer listens, marks errors, subtracts the number of bars in each exercise from a maximum score for the exercise, and totals the exercise scores. While a teacher may know that a student plays at a level far above the first exercises, s/he must still listen to the early exercises to reach a score. Given the large number of students in many band programs, this is probably not the best use of a teacher's time. In administering the test after calibration under the Rasch model, the adjudicator would select an exercise which is an appropriate beginning point from his knowledge of the student. If the student performed 60% or more of the bars correctly, s/he would be asked to play a more difficult exercise. The adjudicator would be expected to use his/her judgment in selecting the next exercise, based on the percent of bars that were played correctly in the initial exercise. Similarly, if the student performed 40% or fewer bars correctly, s/he would be asked to play an easier exercise. This

sequence continues until the student plays an exercise with between 40% and 60% accuracy. This final exercise would correspond to the student's general ability level. From a simple table corresponding to the exercise number, the adjudicator could look up the student's score based on the number of bars played correctly within that exercise.

Furthermore, under classical test theory, which underlies the WFPS, the information gained from the administration of a test is a single number reflecting the number of correct responses. The Rasch model provides this information as well. However, since the items are given difficulty estimates on the same scale as the student ability estimates, a teacher can determine what types of bars a student is and is not likely to perform correctly.

Item Estimates (Thresholds)
all on all (N = 125 L =132)

03/11/97 18:20:29.50



Each X represents 1 students
Some thresholds could not be fitted to the display

BEST COPY AVAILABLE

The variable map shown earlier has been repeated here with example items added. This illustrates what is perhaps the greatest advantage given by the Rasch model. Suppose a new student entered a band program and received an estimate of -1 on the WFPS. The director would know that she would have a 50% chance of correctly playing example item number three and would be more likely than not to play items six and seven correctly and the others incorrectly. Similarly, if a new student scored a 3, she would be likely to play one and two incorrectly, but all the others correctly. A chart could be created from which students could easily read what they need to practice most to achieve the next level.

Again, there are many difficulties in assessing musical performance validly and reliably. The WFPS does not solve, or even address, them all. But the instrument has been shown to have sound psychometric properties and is widely respected. The addition of Rasch scaling improves the efficiency of the administration and makes diagnostic information simple to obtain. More research is of course necessary to calibrate the entire scale. But it appears from this initial foray into Rasch scaling that the Rasch model may provide a promising new assessment tool for both teachers and researchers.

References

Abeles, H., Hoffer,C. & Klotman, R. (1994). Foundations of Music Education. New York: Schirmer Books.

Abeles, Harold F. (1973). A facet-factorial approach to the construction of rating scales to measure complex behaviors. Journal of Educational Measurement, 10 (2), 145-51.

Adams, R, and Khoo, S. (1993). Quest: The Interactive Test Analysis System. Victoria: Australian Council for Educational Research.

Cooksey, J. M. (1977). A facet-factorial approach to rating high school choral music performance. Journal of Research in Music Education, 25, 100-14.

DCamp, Charles B. (1980). Application of the Facet-Factorial Approach to Scale Construction in the Developing of a Rating Scale for High School Band Performance. Unpublished Ph.D. dissertation, University of Iowa, Iowa City, IA.

Kidd, R. L. (1981). The construction and Validation of a Scale of Trombone Performance Skills (Critique). Bulletin of the Council for Research in Music Education, 65, 80-83.

Linacre, J. M. (1989). Many-faceted Rasch Measurement. Chicago: MESA Press.

Massel, P. (1978). The influence of voice quality and the visual element on vocal adjudication. Unpublished master's thesis, University of Western Ontario.

Vasil, T. (1973). The effects of systematically varying selected factors on music performance adjudication. Unpublished doctoral dissertation, Univeristy of Connecticut, Storrs.

Watkins, J. G. & Farnum, S. E. (1954). The Watkins-Farnum Performance Scale: Form A. Winona, Minn.: Hal Leonard, 1954.

Watkins, J. G. (1942). Objective measurement of instrumental performance, New York: Teachers' College Bureau of Publications, Columbia University, 1942.

Wright, B. & Stone, M., (1979). Best Test Design: Rasch Measurement. Chicago, MESA Press.